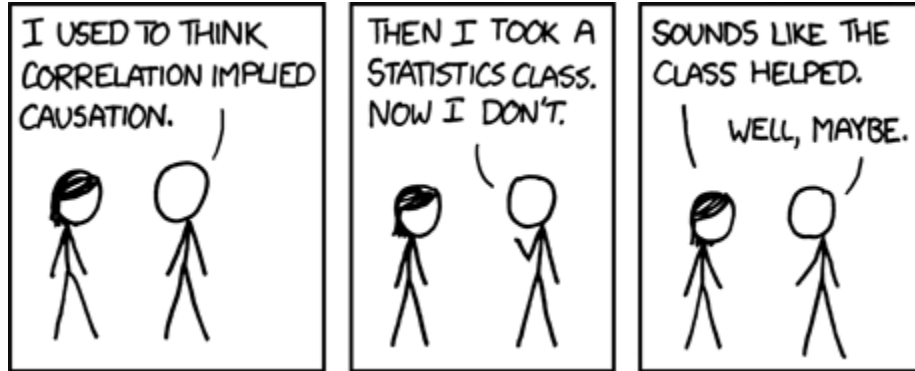
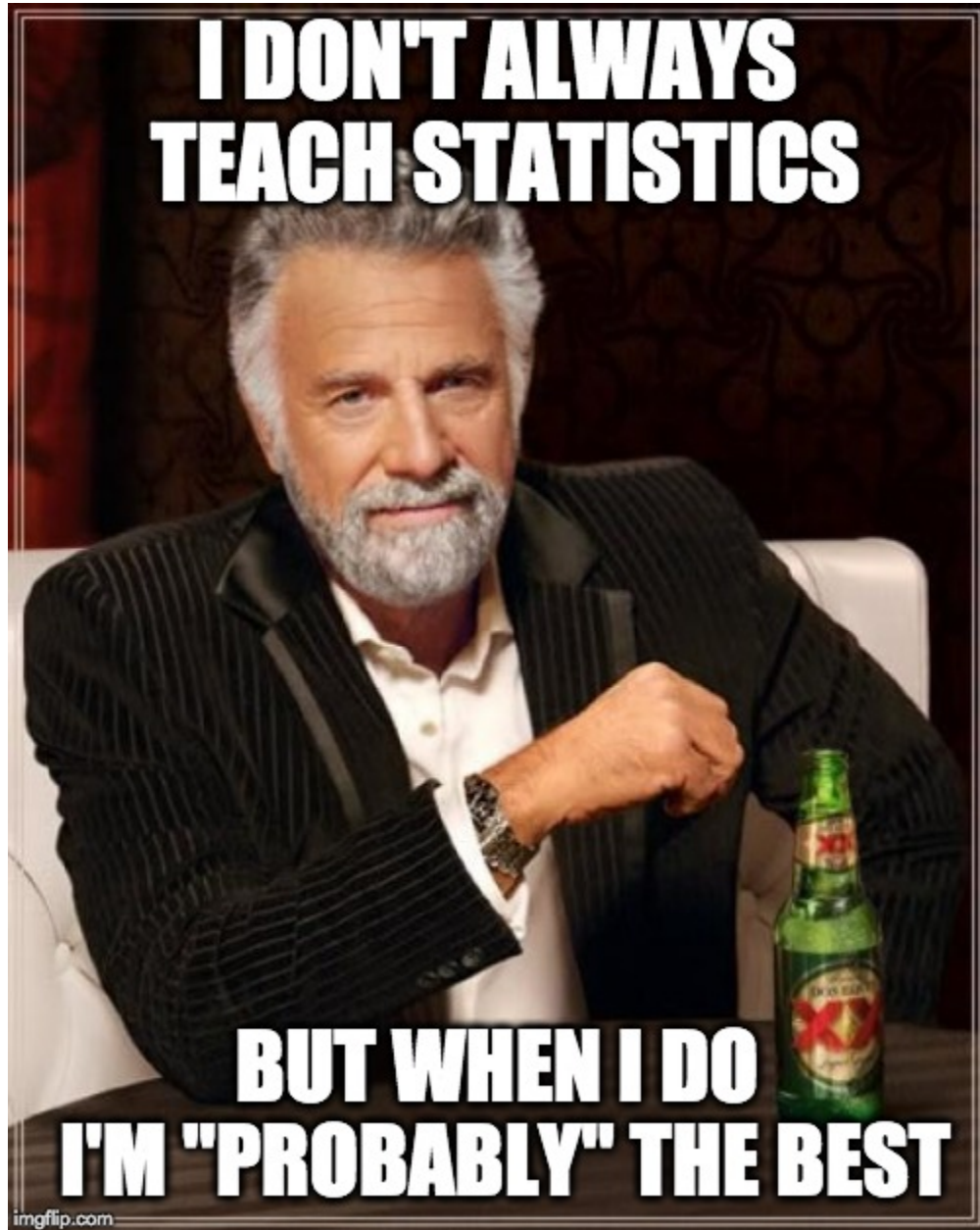


## Unit 1 Live Session - W203, Statistics for Data Science (Summer'20)

### 1. What's statistics? What's this class?



## 2. Instructor Introduction



### 3. Our Students

We will ask you for short introductions on Slack. Question: What does a *data scientist* look like?

### 4. Weekly Workflow

A typical week of the course proceeds as follows:

1. Before live session:
  - A. Watch realated async content.
  - B. Complete related required readings.
2. In live session:
  - A. We will build upon the async to test and extend your understanding
3. After live session:
  - A. complete the homework for the unit (due before next live session)

## 5. Important Resources

- [Course Webpage \(https://w203-summer-20.github.io/class\)](https://w203-summer-20.github.io/class)
  - Up-to-date course policies
  - Calendar with weekly deliverables and deadlines
- [ISVC \(https://learn.datascience.berkeley.edu/login\)](https://learn.datascience.berkeley.edu/login)
  - Platform with course videos
  - Submit your homework/labs here
    - One output file (either PDF or HTML)
    - One source file (IPYNB or RMD or R)
- [Github Org: w203-summer-20 \(https://github.com/w203-summer-20\)](https://github.com/w203-summer-20)
  - Source of homework, labs, and live session documents
  - Check an invite to our *github.org* to your iSchool email
    - it will allow you to join with any github account you like.
- [Slack #w203\\_mids\\_sum\\_20 \(https://ucbischool.slack.com/archives/C012FARHLJE\)](https://ucbischool.slack.com/archives/C012FARHLJE)
  - Our forum for questions about content, announcements, etc.

## 6. Homework (HW)

- Almost weekly.
- Access HW1 via [Github.Org \(https://github.com/w203-summer-20\)](https://github.com/w203-summer-20).
  - Open the repository `unit\_1\_hw`.
  - If you are unfamiliar with *GitHub*, don't worry about the details right now. Find the green button to download a zip file containing all documents. \*You can then open the notebook using Jupyter.
- Submit HW1 via `ISVC>Assessments`
  - There is a page there to upload your solution file

## 7. How to Succeed in this Class

Here are some helpful strategies:

1. Get to know your readings, esp. *Devore* and *Wooldridge* textbooks
  - A. Some questions may come directly from readings
2. Do your best on assignments
3. Strategize about the HW with friends as much as you want, but submitted work must be your individual work.
  - A. Note: do not discuss quizzes/labs with anyone else unless otherwise specified
4. Try each problem on your own for at least a little bit.
  - A. We don't ask plug-and-chug problems in this class.
  - B. 99% of your time will be spent connecting words in English to mathematical objects.
  - C. It's normal not to know what formula to apply - once you do, you're almost done.
  - D. If lost:
    - a. review similar examples in textbook and async videos
    - b. write down relevant definitions
    - c. draw a relevant picture/diagram
    - d. express the unknowns in terms of known variables
  - E. If stuck on a problem for 30 min, talk to a classmate, come to OH, or make a post on class forum
    - a. When posting describe your approach, but do not give out solutions (whether correct or not)
    - b. Give other students a joy of solving the problem on their own
5. Form study groups!
6. Come to office hours (OH)! To all OH!
  - A. Instructors can prepare a specific example, if you send your in advance

## 8. Software

---



You have 3 ways to use R code interpreter:

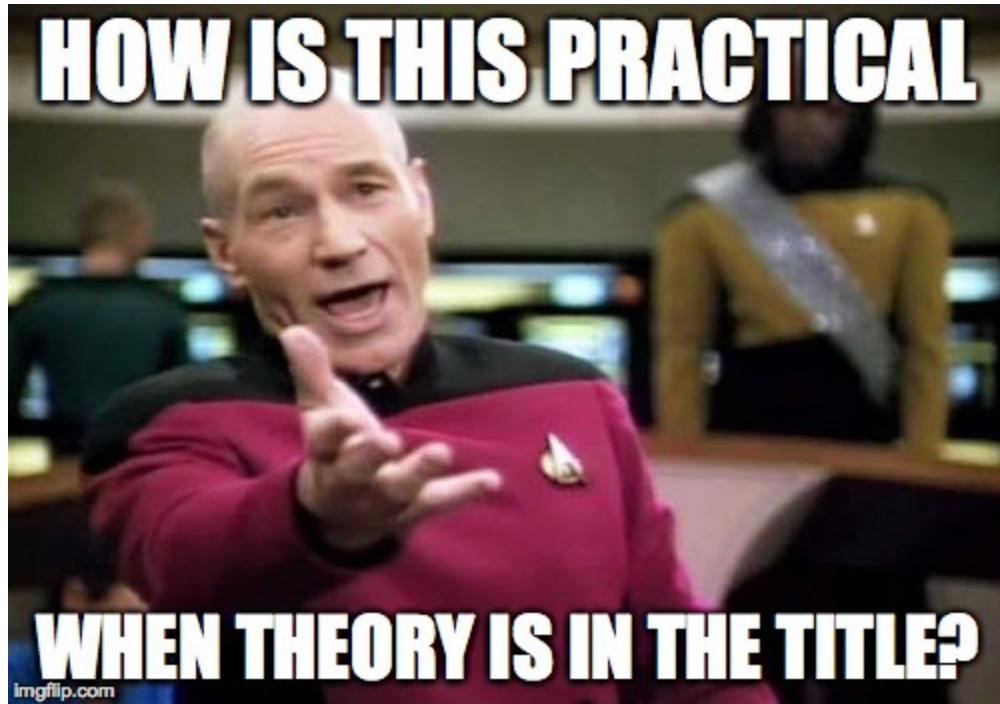
1. [R project](https://cloud.r-project.org) (<https://cloud.r-project.org>) and [RStudio](https://www.rstudio.com/products/rstudio/download) (<https://www.rstudio.com/products/rstudio/download>)
  - A. [Video tutorial](https://www.youtube.com/watch?v=9-RrkJQQYqY) (<https://www.youtube.com/watch?v=9-RrkJQQYqY>)
2. [Anaconda, Jupyter Notebooks](https://www.youtube.com/watch?v=jZ952vChhul) (<https://www.youtube.com/watch?v=jZ952vChhul>) with an R kernel:
  - A. [Video tutorial for Anaconda](https://www.youtube.com/watch?v=YJC6ldl3hWk) (<https://www.youtube.com/watch?v=YJC6ldl3hWk>)
    - a. It contains Jupyter Server and Notebook
  - B. [Video tutorial for R Kernel](https://www.youtube.com/watch?v=SXBxKe8sK6l) (<https://www.youtube.com/watch?v=SXBxKe8sK6l>)
    - a. Just run a command `conda install -c r r-essentials` in [CLI](https://en.wikipedia.org/wiki/Command-line_interface) ([https://en.wikipedia.org/wiki/Command-line\\_interface](https://en.wikipedia.org/wiki/Command-line_interface)) of your operating system (OS)
3. Web-based RStudio using Berkeley's [datahub](https://r.datahub.berkeley.edu) (<https://r.datahub.berkeley.edu>)
  - A. Convenient, if you can't install software on your PC

The first few weeks are heavy in math. Turn in as a single `.pdf` file! You can handwrite/scan or type up (preferred) your solutions with these tools:

1. [LyX](https://www.lyx.org/) (<https://www.lyx.org/>): Visual LaTeX. Excellent tool
2. Microsoft Word has a great formula editor
3. Google Doc has a reasonable formula editor
4. Any Latex editor and turn in the typesetted `.pdf`

LaTeX is painful at first due to errors from failed compilations, but is useful in the long run. We will be able to use the syntax for communicating equations in the notes and chat pods.

## 9. Practical Probability Theory



### Exercise: Sample Space

For each of the following experiments:

1. Give example of an event
2. Define a state space  $\Omega$
3. How big is  $\Omega$  (what is its cardinality, i.e. how many elements does it have?)























































Exercise: Rolling a 6 sided die

Solution: 1. Event is a number that turns up with a throw. For example, "event of three showing up" as displayed. 1.  $\Omega = \{1,2,3,4,5,6\}$  1.  $|\Omega| = 6$

Consider a standard 52-card deck ([https://en.wikipedia.org/wiki/Standard\\_52-card\\_deck](https://en.wikipedia.org/wiki/Standard_52-card_deck)) with 13 ranks and 4 suits. Image source: [Wikipedia \(https://en.wikipedia.org/wiki/Standard\\_52-card\\_deck\)](https://en.wikipedia.org/wiki/Standard_52-card_deck)

Example set of 52 playing cards; 13 of each suit clubs, diamonds, hearts, and spades

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													



### Exercise: Drawing a 5 Card Poker Hand

1. Drawing a particular combination of cards (such as the one shown) is an event in  $\Omega$  1.  $\Omega$  includes all possible combinations of 5 cards from a set of 52 cards. 1.

$$|\Omega| = C_5^{52} = \frac{52!}{(52 - 5)! \cdot 5!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2} = 2,598,960$$



```
In [1]: cat(paste('Omega size: ', choose(52, 5), '\n'))

Ranks = paste(c('A', seq(2,10), 'J', 'Q', 'K'))
Suits = c('c','d','h','s')

RS_comb <- expand.grid(Ranks, Suits) # rank-suit combinations
Cards = paste0(c(RS_comb)$Var1, c(RS_comb)$Var2)

cat(paste('Number of cards: ', length(Cards), '\n'))
cat(paste('All 52 cards:', paste(Cards, collapse=','), '\n'))
cat('Some of the 5 card hands:\n')
combn(Cards, 5)[,1:20]
```

Omega size: 2598960

Number of cards: 52

All 52 cards: Ac,2c,3c,4c,5c,6c,7c,8c,9c,10c,Jc,Qc,Kc,Ad,2d,3d,4d,5d,6d,7d,8d,9d,10d,Jd,Qd,Kd,Ah,2h,3h,4h,5h,6h,7h,8h,9h,10h,Jh,Qh,Kh,As,2s,3s,4s,5s,6s,7s,8s,9s,10s,Js,Qs,Ks

Some of the 5 card hands:



Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac
2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c
3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c	3c
4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c	4c
5c	6c	7c	8c	9c	10c	Jc	Qc	Kc	Ad	2d	3d	4d	5d	6d	7d	8d	9d	10d	Jd



### Exercise: Choosing 1000 U.S. citizens for a survey

1. Any subset of 1000 individuals from the U.S. citizens population (~325M) 1.  $\Omega$  includes all possible combinations of 1000 unique individuals from 325M. 1.  $|\Omega| =$  a number so large that R shows `inf`

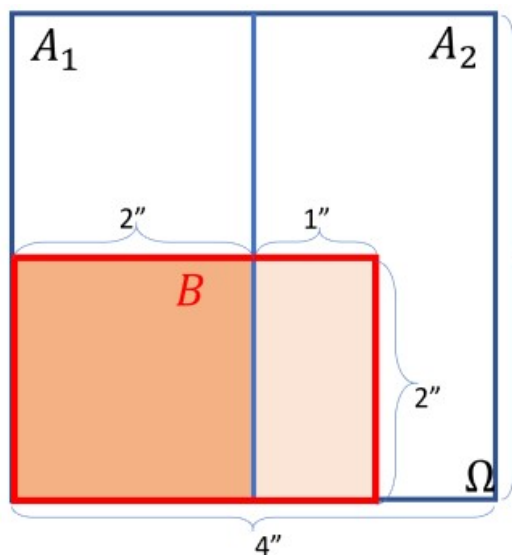
```
In [6]: USPopulation = 325000000
choose(USPopulation,10)    # Number of ways to choose 10 individuals (in any order) from 325M
choose(USPopulation,100)   # Number of ways to choose 100 individuals (in any order) from 325M
choose(USPopulation,1000)  # Number of ways to choose 1000 individuals (in any order) from 325M
```

180.888944691403

Inf

Inf

### Exercise: Compute with LTP



Law of Total Probability.  
Consider sample space  $\Omega$  as  $2 \times 2$  square of points partitioned into halves by  $A_1, A_2$ . Also  $B$  is a subset of events as shown in figure. All events are equi-weighted.

4" The shapes are simple and we can directly compute the  $\mathbb{P}B = \frac{2 \cdot 3}{4 \cdot 4} = \frac{3}{8}$ , but sometimes it is easier to apply the Law of Total Probability:

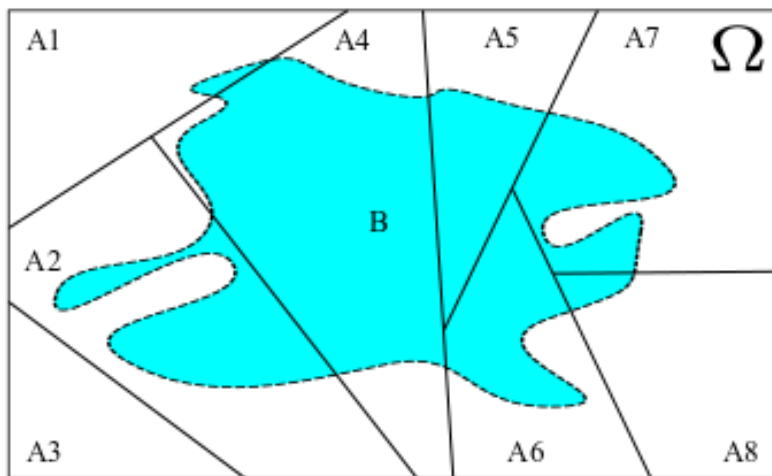
$$\mathbb{P}B = \mathbb{P}[B|A_1] \cdot \mathbb{P}A_1 + \mathbb{P}[B|A_2] \cdot \mathbb{P}A_2$$

What are the probabilities in the summation?

Recall (Devore,p76):  $\mathbb{P}[B|A_1] = \frac{\mathbb{P}[B \cap A_1]}{\mathbb{P}A_1} = \frac{2 \cdot 2}{2 \cdot 4} = 4/8$  and so on.  

$$\mathbb{P}B = 4/8 \cdot 1/2 + 2/8 \cdot 1/2 = 1/4 + 1/8 = 3/8$$
  
 Thankfully, this gives us the same answer ;)

## Exercise: Explain LTP



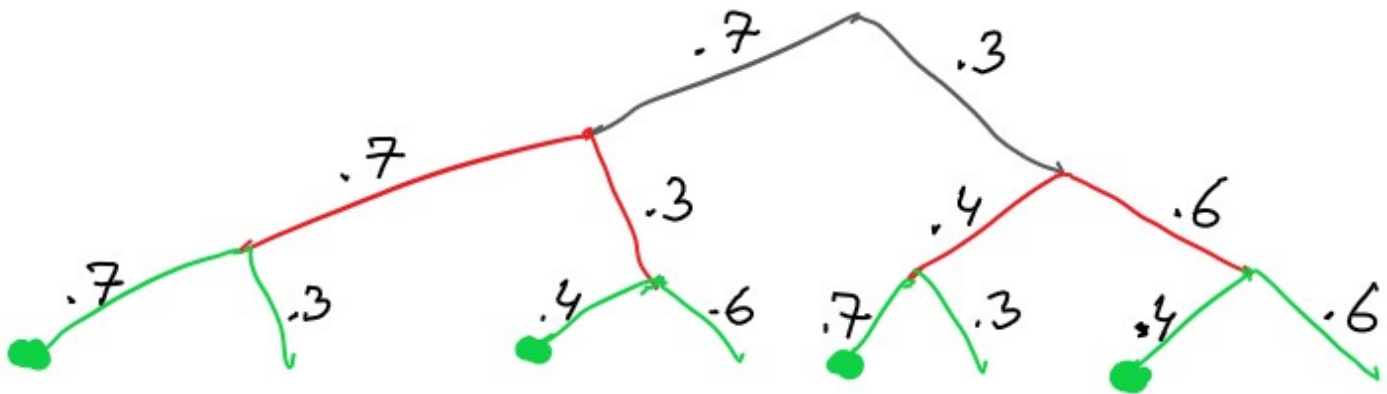
Consider a partition of  $\Omega$  state space 1. How would an  $A_i$ 's look? 1. How would a disjoint, but not an exhaustive set of  $A_i$ 's look? 1. Explain what the Law of Total Probabilities (LTP) means intuitively.

$$P(B) = \sum_{i=1}^N P(A_i \cap B) = \sum_{i=1}^N P(B|A_i)P(A_i)$$

## Exercise: Conditional Probabilities

Suppose you're taking a statistics class and that in each week you are either caught up or behind on the readings. It is difficult to incorporate a new class in to your schedule so the probability that you will be caught up in Week 1, having viewed all async and completed the pre class exercises before the live session, is 0.7. From then on, if you are caught up in a given week, the probability that you will be caught up in the next week is 0.7. If you are behind in a given week, the probability that you will be caught up in the next week is 0.4. What is the probability that you are caught up in week 3?

You can use a tree diagram or a formula (which one?) to solve the problem.



The green thick leaf nodes indicate successful catch up in week 3. Now we simply compute the probabilities. Denote probability of successfully catching up in week  $i$  as  $\mathbb{P}S_i$  and the probability failing to catch up is complementary, i.e.  $\mathbb{P}[\neg S_1] = 0.3$ . In week 1, you have: 1.  $\mathbb{P}[S_1] = 0.7$  1.  $\mathbb{P}[\neg S_1] = 1 - \mathbb{P}[S_1] = 0.3$  1. In week 2, the catch up depends on the how you completed the prior week: 1.  $\mathbb{P}[S_2|S_1] = 0.7$  1.  $\mathbb{P}[\neg S_2|S_1] = 1 - \mathbb{P}[S_2|S_1] = 0.3$  1.  $\mathbb{P}[S_2|\neg S_1] = 0.4$  1.  $\mathbb{P}[\neg S_2|\neg S_1] = 1 - \mathbb{P}[S_2|\neg S_1] = 0.6$  1. In week 3, the catch up depends on the how you completed the prior week: ...

In [3]: `.7*.7*.7+.7*.3*.4+.3*.4*.7+.3*.6*.4`

0.583

## Exercise

A test for certain disease is assumed to be correct 95% of the time: if a person has the disease the test will give a positive result with probability 0.95. If a person does not have disease the test will give a negative result with probability 0.95. A random person drawn from a certain population has a probability 0.001 of having the disease. Given that a person drawn at random just tested positive, what is the probability that they have the disease? Setup:

Similar to above, we can either build a tree or state all probabilities symbolically. Let  $D$  be the event of a person having the disease, "+" is the event of the test showing positive, and "-" is the event of the test showing negative. Then we need to find  $\mathbb{P}[D|+]$

**We have:** 1. Test is always correct: 1. True positive rate (TPR):  $\mathbb{P}[+|D] = 0.95$  1. So, false positive rate (FPR):  $\mathbb{P}[+|!D] = 1 - \mathbb{P}[+|D] = 0.05$  1. True negative rate (TNR):  $\mathbb{P}[-|!D] = 0.95$  1. So, false negative rate (FNR):  $\mathbb{P}[-|D] = 1 - \mathbb{P}[-|!D] = 0.05$  1. Population infection rate:  $\mathbb{P}[D] = .001$  1. Then  $\mathbb{P}[!D] = .999$  Let's put Bayes' formula to use that might give us some known components:

$$\mathbb{P}[D|+] = \frac{\mathbb{P}[+|D] \cdot \mathbb{P}[D]}{\mathbb{P}[+]}$$

**We know all quantities on the r.h.s., except  $\mathbb{P}[+]$ , which we can find with the LTP:**

$$\mathbb{P}[+] = \mathbb{P}[+|D] \cdot \mathbb{P}[D] + \mathbb{P}[+|!D] \mathbb{P}[!D]$$

**Then plug into your favorite calculator (R-hint-hint) to derive 0.01866.**

## 10. Reminders

Before next week:

1. Install Jupyter with an R kernel (or make sure you can use [r.datahub.berkeley.edu](https://r.datahub.berkeley.edu))
2. Complete the Unit 1 homework located in our Github.org (due before next live session).
3. Watch all unit 2 async content

In [ ]: